ORIGINAL

NATIONAL UNIVERSITY OF SINGAPORE

EXAMINATION FOR

(Semester I : 2017/2018)

EE5907 – PATTERN RECOGNITION

Nov 2017 – Time Allowed: 2.5 Hours

INSTRUCTIONS TO CANDIDATES

1. This paper contains **FOUR (4)** questions and comprises **FIVE (5)** printed pages.

2. All questions are compulsory. Answer **ALL** questions. The examination paper takes **ONE HUNDRED (100)** marks in total.

3. This is a **CLOSED BOOK** examination. One A4-size formula sheet is allowed.

4. Programmable calculators are not allowed.

Q1 **(25 marks)**. Subquestions (a), (b) and (c) can be answered independently.

(a) Consider training data $x_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, x_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, x_3 = \begin{pmatrix} 1.5 \\ 0 \end{pmatrix}, x_4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$ with corresponding class labels $y_1 = 0, y_2 = 0, y_3 = 1, y_4 = 1$. What is the 3-NN estimate of the class label posterior probabilities of datapoints $x_5 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, where the distance metric used is the Euclidean distance? What is the MAP classification of data points x_5? Repeat the above with the Manhattan distance metric.

(6 marks)

(b) Let $D = \{x_1, \ldots, x_N\}$ be independent samples of a Poisson distribution with unknown parameter λ. In other words, $p(x_n|\lambda) = e^{-\lambda}\frac{\lambda^{x_n}}{x_n!}$. Derive the maximum likelihood (ML) estimate of λ.

(7 marks)

(c) Consider a binary classification problem of predicting binary class y from features x. The cost of correct prediction is \$0. There is a \$7 cost associated with predicting class 0 when the true class is 1. There is a \$3 cost associated with predicting class 1 when the true class is 0. Suppose the cost of asking a human to perform the manual classification is \$2. Therefore for a particular x, there are three possible decisions: (1) decision α_0 predicts y to be 0, (2) decision α_1 predicts y to be 1 and (3) decision α_h requires a human to perform the manual classification. Let $p_1 = p(y = 1|x)$

(i) Assume the human is 100% accurate. What is the general decision rule (as a function of p_1) in order to minimize expected loss?

(6 marks)

(ii) Assume the human is only 90% accurate. Assume that when the human is wrong, the correct class is equally likely to be class 0 or class 1. What is the general decision rule (as a function of p_1) in order to minimize expected loss?

(6 marks)

Q2 **(25 marks)**. Subquestions (a) and (b) can be answered independently.

(a) Consider a 2-class naïve Bayes classifier with one binary feature and one Gaussian feature. More specifically, class label y follows a categorical distribution parametrized by π, i.e., $p(y = c) = \pi_c$. The first feature x_1 is binary and follows a Bernoulli distribution: $p(x_1|y = c) = \text{Bernoulli}(x_1|\theta_c)$. The second feature x_2 is a univariate Gaussian: $p(x_2|y = c) = \mathcal{N}(x_2|\mu_c, \sigma_c^2)$. Let $\pi = [0.75\ 0.25]$, $\theta = [0.4\ 0.5]$, $\mu = [-1\ 0]$ and $\sigma^2 = [1\ 1]$.

(i) Compute $p(y|x_1 = 0)$. Note that result is a vector of length 2 that sums to 1.
(6 marks)

(ii) Compute $p(y|x_2 = 0)$. Note that result is a vector of length 2 that sums to 1.
(8 marks)

(iii)Compute $p(y|x_1 = 0, x_2 = 0)$. Note that result is a vector of length 2 that sums to 1.
(6 marks)

(b) Suppose you are a company trying to decide how much quantity Q of some product (e.g., newspapers) to produce to maximize your profit z. The optimal amount will depend on how much demand D you think there is for your product, as well as its per unit cost C and selling price P. Suppose D is unknown but has probability distribution function $f(D)$ and cumulative distribution function $F(D)$. We can evaluate the expected profit by considering two cases: if $D > Q$, then we sell all Q items, and make profit $z = (P - C)Q$; but if $D < Q$, we only sell D items, at profit $z = (P - C)D$, but have wasted $C(Q - D)$ on the unsold items. So the expected profit for quantity Q is given by

$$E_Q(z) = \int_Q^{\infty} (P - C)Qf(D)dD + \int_0^Q (P - C)Df(D)dD - \int_0^Q C(Q - D)f(D)dD$$

Simplify this expression, and then take derivatives with respect to Q to derive an expression (i.e., mathematical equation) that the optimal quantity Q^* should satisfy. Your final expression should be in terms of $F(Q^*)$, P and C. Note that you do not need to solve for the optimal Q^* (which is not possible since F is not given to you).

(5 marks)

Q3 **(25 marks)**. Given labelled d-dimensional training vectors $x \in R^d$ from C classes, with $n_i$vectors from class c_i for $i = 1, \ldots, C$ and $\sum_{i=1}^{C} n_i = n$. Linear discriminative analysis (LDA) projection direction, $W \in R^{d \times p}$, is obtained by maximizing

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|}$$

where $S_B \in R^{d \times d}$ and $S_w \in R^{d \times d}$ are the between class scatter and within class scatter respectively.

(a) Derive the expression for the optimal projection direction W^* for $C = 2$.

(10 marks)

(b) Explain the difference between LDA and Marginal Fisher Analysis (MFA). For what kind of data distribution, MFA can perform better than LDA? Please give one example of such data distribution and explain the reason.

(8 marks)

(c) For Principal Component Analysis (PCA), the objective to maximize can be expressed as $Q(W) = \text{trace}(W^T S_T W)$ with $S_T = S_B + S_W$ and $W^T W = I_p$. Based on $J(W)$ and $Q(W)$, explain why PCA is generally worse than LDA when the dimension-reduced features are used for classification purpose.

(7 marks)

Q4 **(25 marks)**. The following sub-problems are independent.

(a) Given a data matrix X, describe how the Nonnegative Matrix Factorization (NMF) algorithm finds the basis matrix W and coefficient matrix H. Please list the objective function of NMF, its optimization steps and derivation of the updating rules for W and H.

(8 marks)

(b) The sensitivity to initial centroids is a key issue for k-means, please list the popular solutions to selection on proper centroids.

(3 marks)

(c) What are the general differences between generative models and discriminative models? Please give three application examples where the discriminative model is more preferred than the generative model.

(5 marks)

(d) Follow Kuhn-Tucker theorem to convert the primal constrained optimization problem in two class Support Vector Machine (SVM) into a dual, unconstrained one. The primal optimization problem is given by:

$$\begin{aligned} &\min \ \tfrac{1}{2}\|w\|_2^2 \\ &\text{s.t. } y_i w^T x_i \geq 1, i = 1, \ldots, n \end{aligned}$$

where (x_i, y_i) is a training data, $x_i \in R^d$ is a feature and $y_i \in \{-1, +1\}$ is the corresponding label. Derive the following dual form:

$$\max \ -\frac{1}{2}\sum_{j,k=1}^{n} \alpha_j \alpha_k y_j y_k (x_j^T x_k) + \sum_{j=1}^{n} \alpha_j.$$

(9 marks)

END OF PAPER